Rodney Shaghoulian – shaghou2 – (3 Units)

2.1 Text Document Classification

The backbone of the Multinomial Naive Bayes solver is the **Data** class. The goal of this class was to save all sorts of information, in various ways, from reading the text files. This made coding the Multinomial Naive Bayes and Bernoulli Naive Bayes solver much simpler to implement.

```
public class Data {
      int numEntries;
      int numSpamEntries;
      int numNotSpamEntries;

      HashMap<String, Integer> wordsNoCategory;
      HashMap<String, Integer> spamWords;
      HashMap<String, Integer> notSpamWords;

      int totalWords;
      int totalSpamWords;
      int totalNotSpamWords;

      int uniqueWords;
      int uniqueSpamWords;
      int uniqueNotSpamWords;

      ArrayList<HashMap<String, Integer>> entries;
      ArrayList<HashMap<String, Integer>> spamEntries;
      ArrayList<HashMap<String, Integer>> notSpamEntries;
      ArrayList<Boolean> isSpam;

      // Methods ommited
}
```

The **Data** class prepares useful information in various formats to be used by the **MAP** class. The **MAP** class performs a **maximum a posteriori** classification of the text documents, using:

- Multinomial Naive Bayes
- Bernoulli Naive Bayes

```
public class MAP {
      /* Priors, Likelihoods, Posteriors */
      double priorSpam;
      double priorNotSpam;
      HashMap<String, Double> likelihoodsSpam;
      HashMap<String, Double> likelihoodsNotSpam;
      double [] posteriorSpam;
      double [] posteriorNotSpam;

      // Methods omitted (including methods to calculate priors/likelihoods/posteriors)
}
```

We use “k = 1” for Laplace Smoothing

We will see that the **Multinomial Naive Bayes** model provies better results than the **Bernoulli Naive Bayes** model.

Multinomial Naive Bayes – SPAM DETECTION

Overall Prediction Accuracy = 0.969

*** Classification Rates ***

Spam Prediction Accuracy = 0.985
Not Spam Prediction Accuracy = 0.954

*** 20 most common SPAM words ***

s
our
free
please
one
mail
over
day
us
want
address
send
com
http
information
remove
product
order
offer
need

*** 20 most common NOT SPAM words ***

language
s
university
linguistic
information
email
please
conference
e
research
include
one
english
paper
address
workshop
edu
papers
fax
http

*** Confusion Matrix ***

The following is the Confusion Matrix where each entry in row “r” and column “c” is the percentage of test images from class “r” that are classified as class “c”

	Spam	**Not Spam**
Spam	0.985	0.015
Not Spam	0.046	0.954

Multinomial Naive Bayes – MOVIE REVIEWS

Overall Prediction Accuracy = 0.761

***** Classification Rates *****

Positive Review Prediction Accuracy = 0.766
Negative Review Prediction Accuracy = 0.756

*** 20 most common POSITIVE REVIEW words ***

film
movie
--
one
like
story
good
comedy
way
even
time
best
much
performances
make
funny
us
life
makes
characters

*** 20 most common NEGATIVE REVIEW words ***

movie
film
like
one
--
bad
story
much
time
even
good
characters
little
would
comedy
never
nothing
plot
makes
make

***** Confusion Matrix *****

The following is the Confusion Matrix where each entry in row “r” and column “c” is the percentage of test images from class “r” that are classified as class “c”

	Positive Review	**Negative Review**
Positive Review	0.766	0.234
Negative Review	0.244	0.756

Bernoulli Naive Bayes – SPAM DETECTION

Overall Prediction Accuracy = 0.958

***** Classification Rates *****

Spam Prediction Accuracy = 0.954
Not Spam Prediction Accuracy = 0.962

***** 20 most common SPAM words *****

our
please
free
s
one
us
day
http
remove
mail
com
here
information
want
over
need
offer
receive
list
address

***** 20 most common NOT SPAM words *****

language
s
university
information
linguistic
http
email
please
e
fax
follow
include
one
english
call
research
www
interest
address
word

***** Confusion Matrix *****

The following is the Confusion Matrix where each entry in row “r” and column “c” is the percentage of test images from class “r” that are classified as class “c”

	Spam	**Not Spam**
Spam	0.954	0.046
Not Spam	0.038	0.962

Bernoulli Naive Bayes – MOVIE REVIEWS

Overall Prediction Accuracy = 0.755

*** Classification Rates ***

Positive Review Prediction Accuracy = 0.742
Negative Review Prediction Accuracy = 0.768

* 20 most common POSITIVE REVIEW words *

```
film
movie
one
like
--
story
comedy
way
even
good
best
time
much
performances
funny
makes
make
life
characters
work
```

* 20 most common NEGATIVE REVIEW words *

```
movie
film
like
one
story
much
--
bad
time
even
characters
little
good
would
comedy
nothing
plot
makes
never
make
```

*** Confusion Matrix ***

The following is the Confusion Matrix where each entry in row “r” and column “c” is the percentage of test images from class “r” that are classified as class “c”

	Positive Review	**Negative Review**
Positive Review	0.742	0.258
Negative Review	0.232	0.768

2.2 NewsGroup Dataset (Extra Credit)

Multinomial Naive Bayes

For Laplace smoothing, we use “k = 1”.

For the newsgroup dataset we are going to use the following abbreviations

Category Number	Category Name
0	sci.space
1	comp.sys.ibm.pc.hardware
2	rec.sport.baseball
3	comp.windows.x
4	talk.politics.misc
5	misc.forsale
6	rec.sport.hockey
7	comp.graphics

Overall Prediction Accuracy = 0.920

***** Classification Rates *****

```
categoryPrediction[0] = 0.971
categoryPrediction[1] = 0.848
categoryPrediction[2] = 0.972
categoryPrediction[3] = 0.893
categoryPrediction[4] = 0.979
categoryPrediction[5] = 0.500
categoryPrediction[6] = 1.000
categoryPrediction[7] = 0.828
```

***** Confusion Matrix *****

```
     0      1      2      3      4      5      6      7
0: 0.971  0.000  0.000  0.000  0.029  0.000  0.000  0.000
1: 0.000  0.848  0.000  0.091  0.030  0.000  0.000  0.030
2: 0.000  0.000  0.972  0.000  0.000  0.000  0.028  0.000
3: 0.000  0.000  0.000  0.893  0.036  0.000  0.000  0.071
4: 0.021  0.000  0.000  0.000  0.979  0.000  0.000  0.000
5: 0.000  0.400  0.000  0.000  0.100  0.500  0.000  0.000
6: 0.000  0.000  0.000  0.000  0.000  0.000  1.000  0.000
7: 0.034  0.034  0.000  0.069  0.034  0.000  0.000  0.828
```

```
*** 20 most common words. Category: 0 ***
nt
would
space
subject
one
like
writes
could
us
article
time
edu
also
orbit
much
earth
get
nasa
people
launch
```

```
*** 20 most common words. Category: 1 ***
nt
one
drive
subject
would
card
ide
get
use
m
bus
system
controller
two
know
also
edu
like
writes
article
```

```
*** 20 most common words. Category: 2 ***
nt
edu
would
writes
year
one
subject
last
article
game
think
good
like
team
players
better
baseball
well
time
get
```

```
*** 20 most common words. Category: 3 ***
x
nt
subject
use
window
like
using
one
would
get
code
windows
writes
edu
server
running
run
problem
motif
article
```

```
*** 20 most common words. Category: 4 ***
nt
would
people
writes
article
one
edu
like
subject
government
us
think
even
make
could
right
m
get
new
much
```

```
*** 20 most common words. Category: 5 ***
edu
new
subject
sale
shipping
one
price
comics
nt
email
please
card
hulk
spiderman
get
condition
cover
offer
issue
original
```

```
*** 20 most common words. Category: 6 ***
nt
team
game
hockey
would
subject
one
think
year
nhl
first
like
play
go
games
get
writes
players
edu
time
```

```
*** 20 most common words. Category: 7 ***
nt
one
subject
would
graphics
edu
like
use
computer
also
writes
know
x
information
article
line
get
bit
need
new
```

Bernoulli Naive Bayes

For Laplace smoothing, we use “k = 0.01” since “k = 1” gave a much lower prediction accuracy.

For the newsgroup dataset we are going to use the following abbreviations

Category Number	Category Name
0	sci.space
1	comp.sys.ibm.pc.hardware
2	rec.sport.baseball
3	comp.windows.x
4	talk.politics.misc
5	misc.forsale
6	rec.sport.hockey
7	comp.graphics

```
Overall Prediction Accuracy = 0.920
```

```
*** Classification Rates ***

categoryPrediction[0] = 1.000
categoryPrediction[1] = 0.939
categoryPrediction[2] = 0.972
categoryPrediction[3] = 0.929
categoryPrediction[4] = 0.979
categoryPrediction[5] = 0.900
categoryPrediction[6] = 1.000
categoryPrediction[7] = 0.517
```

```
*** Confusion Matrix ***

     0      1      2      3      4      5      6      7
0: 1.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
1: 0.000  0.939  0.000  0.061  0.000  0.000  0.000  0.000
2: 0.000  0.000  0.972  0.000  0.000  0.000  0.028  0.000
3: 0.000  0.036  0.036  0.929  0.000  0.000  0.000  0.000
4: 0.021  0.000  0.000  0.000  0.979  0.000  0.000  0.000
5: 0.000  0.100  0.000  0.000  0.000  0.900  0.000  0.000
6: 0.000  0.000  0.000  0.000  0.000  0.000  1.000  0.000
7: 0.034  0.310  0.000  0.103  0.000  0.034  0.000  0.517
```

```
*** 20 most common words. Category: 0 ***
subject
would
nt
writes
article
space
one
like
could
also
get
think
us
time
much
new
way
see
m
edu
```

```
*** 20 most common words. Category: 1 ***
subject
nt
one
would
writes
use
know
get
article
card
like
also
two
m
edu
system
work
drive
problem
could
```

```
*** 20 most common words. Category: 2 ***
subject
writes
nt
article
edu
would
one
last
like
year
baseball
think
good
get
m
time
know
game
first
team
```

```
*** 20 most common words. Category: 3 ***
subject
nt
x
use
writes
article
get
using
like
would
one
window
problem
know
also
code
m
help
email
need
```

```
*** 20 most common words. Category: 4 ***
subject
writes
article
nt
people
would
one
like
edu
even
us
think
m
could
get
make
government
know
much
time
```

```
*** 20 most common words. Category: 5 ***
subject
sale
edu
shipping
new
please
email
price
nt
one
get
condition
like
used
good
want
list
use
know
etc
```

```
*** 20 most common words. Category: 6 ***
subject
nt
team
hockey
writes
game
one
would
article
like
first
think
play
go
get
nhl
year
time
games
last
```

```
*** 20 most common words. Category: 7 ***
subject
nt
one
writes
would
article
like
also
know
graphics
get
use
edu
need
computer
could
think
m
two
well
```